APPLICATION

FOR

UNITED STATES LETTERS PATENT

TITLE: IMPROVED SPEECH MODEL AND ANALYSIS, SYNTHESIS, AND QUANTIZATION METHODS

APPLICANT: DANIEL W. GRIFFIN AND JOHN C. HARDWICK

Fish & Richardson P.C.
601 Thirteenth Street, NW
Washington, DC 20005
Tel.: (202) 783-5070
Fax: (202) 783-2331

Improved Speech Model and Analysis, Synthesis, and Quantization Methods

Background

The invention relates to an improved model of speech or acoustic signals and methods for estimating the improved model parameters and synthesizing signals from these parameters.

Speech models together with speech analysis and synthesis methods are widely used in applications such as telecommunications, speech recognition, speaker identification, and speech synthesis. Vocoders are a class of speech analysis/synthesis systems based on an underlying model of speech. Vocoders have been extensively used in practice. Examples of vocoders include linear prediction vocoders, homomorphic vocoders, channel vocoders, sinusoidal transform coders (STC), multiband excitation (MBE) vocoders, improved multiband excitation (IMBE^{TM}), and advanced multiband excitation vocoders (AMBE^{TM}).

Vocoders typically model speech over a short interval of time as the response of a system excited by some form of excitation. Typically, an input signal $s_0(n)$ is obtained by sampling an analog input signal. For applications such as speech coding or speech recognition, the sampling rate ranges typically between 6 kHz and 16 kHz. The method works well for any sampling rate with corresponding changes in the associated parameters. To focus on a short interval centered at time t, the input signal $s_0(n)$ is typically multiplied by a window $w(t,n)$ centered at time t to obtain a windowed signal $s(t,n)$. The window used is typically a Hamming window or Kaiser window and can be constant as a function of t so that $w(t,n) = w_0(n-t)$ or can have characteristics which change as a function of t. The length of the window $w(t,n)$ typically ranges between 5 ms and 40 ms. The windowed signal $s(t,n)$ is typically computed at center times of $t_0, t_1, ...t_m, t_{m+1}, ...$ Typically, the interval between consecutive center times $t_{m+1} - t_m$ approximates the effective length of the window $w(t,n)$ used for these center times. The windowed signal $s(t,n)$ for a particular center time is often referred to as a segment or frame of the input signal.

For each segment of the input signal, system parameters and excitation parameters are determined. The system parameters typically consist of the spectral envelope or the impulse response of the system. The excitation parameters typically consist of a fundamental frequency (or pitch period) and a voiced/unvoiced (V/UV) parameter which indicates whether the input signal has pitch (or indicates the degree to which the input signal has pitch). For vocoders such as

MBE, IMBE, and AMBE, the input signal is divided into frequency bands and the excitation parameters may also include a V/UV decision for each frequency band. High quality speech reproduction may be provided using a high quality speech model, an accurate estimation of the speech model parameters, and high quality synthesis methods.

When the voiced/unvoiced information consists of a single voiced/unvoiced decision for the entire frequency band, the synthesized speech tends to have a "buzzy" quality especially noticeable in regions of speech which contain mixed voicing or in voiced regions of noisy speech. A number of mixed excitation models have been proposed as potential solutions to the problem of "buzziness" in vocoders. In these models, periodic and noise-like excitations which have either time-invariant or time-varying spectral shapes are mixed.

In excitation models having time-invariant spectral shapes, the excitation signal consists of the sum of a periodic source and a noise source with fixed spectral envelopes. The mixture ratio controls the relative amplitudes of the periodic and noise sources. Examples of such models are described by Itakura and Saito, "Analysis Synthesis Telephony Based upon the Maximum Likelihood Method," *Reports of 6th Int. Cong. Acoust.*, Tokyo, Japan, Paper C-5-5, pp. C17-20, 1968; and Kwon and Goldberg, "An Enhanced LPC Vocoder with No Voiced/Unvoiced Switch," *IEEE Trans. on Acoust., Speech, and Signal Processing*, vol. ASSP-32, no. 4, pp. 851-858, August 1984. In these excitation models, a white noise source is added to a white periodic source. The mixture ratio between these sources is estimated from the height of the peak of the autocorrelation of the LPC residual.

In excitation models having time-varying spectral shapes, the excitation signal consists of the sum of a periodic source and a noise source with time varying spectral envelope shapes. Examples of such models are decribed by Fujimara, "An Approximation to Voice Aperiodicity," *IEEE Trans. Audio and Electroacoust.*, pp. 68-72, March 1968; Makhoul et al, "A Mixed-Source Excitation Model for Speech Compression and Synthesis," *IEEE Int. Conf. on Acoust. Sp. & Sig. Proc.*, April 1978, pp. 163-166; Kwon and Goldberg, "An Enhanced LPC Vocoder with No Voiced/Unvoiced Switch," *IEEE Trans. on Acoust., Speech, and Signal Processing*, vol. ASSP-32, no. 4, pp. 851-858, August 1984; and Griffin and Lim, "Multiband Excitation Vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 1223-1235, Aug. 1988.

In the excitation model proposed by Fujimara, the excitation spectrum is divided into three fixed frequency bands. A separate cepstral analysis is performed for each frequency band and a voiced/unvoiced decision for each frequency band is made based on the height of the

cepstrum peak as a measure of periodicity.

In the excitation model proposed by Makhoul et al., the excitation signal consists of the sum of a low-pass periodic source and a high-pass noise source. The low-pass periodic source is generated by filtering a white pulse source with a variable cut-off low-pass filter. Similarly, the high-pass noise source was generated by filtering a white noise source with a variable cut-off high-pass filter. The cut-off frequencies for the two filters are equal and are estimated by choosing the highest frequency at which the spectrum is periodic. Periodicity of the spectrum is determined by examining the separation between consecutive peaks and determining whether the separations are the same, within some tolerance level.

In a second excitation model implemented by Kwon and Goldberg, a pulse source is passed through a variable gain low-pass filter and added to itself, and a white noise source is passed through a variable gain high-pass filter and added to itself. The excitation signal is the sum of the resultant pulse and noise sources with the relative amplitudes controlled by a voiced/unvoiced mixture ratio. The filter gains and voiced/unvoiced mixture ratio are estimated from the LPC residual signal with the constraint that the spectral envelope of the resultant excitation signal is flat.

In the multiband excitation model proposed by Griffin and Lim, a frequency dependent voiced/unvoiced mixture function is proposed. This model is restricted to a frequency dependent binary voiced/unvoiced decision for coding purposes. A further restriction of this model divides the spectrum into a finite number of frequency bands with a binary voiced/unvoiced decision for each band. The voiced/unvoiced information is estimated by comparing the speech spectrum to the closest periodic spectrum. When the error is below a threshold, the band is marked voiced, otherwise, the band is marked unvoiced.

The Fourier transform of the windowed signal $s(t,n)$ will be denoted by $S(t,\omega)$ and will be referred to as the signal Short-Time Fourier Transform (STFT). Suppose $s_0(n)$ is a periodic signal with a fundamental frequency ω_0 or pitch period n_0. The parameters ω_0 and n_0 are related to each other by $2\pi/\omega_0 = n_0$. Non-integer values of the pitch period n_0 are often used in practice.

A speech signal $s_0(n)$ can be divided into multiple frequency bands using bandpass filters. Characteristics of these bandpass filters are allowed to change as a function of time and/or frequency. A speech signal can also be divided into multiple bands by applying frequency windows or weightings to the speech signal STFT $S(t,\omega)$.

Summary

In one aspect, generally, methods for synthesizing high quality speech use an improved speech model. The improved speech model is augmented beyond the time and frequency dependent voiced/unvoiced mixture function of the multiband excitation model to allow a mixture of three different signals. In addition to parameters which control the proportion of quasi-periodic and noise-like signals in each frequency band, a parameter is added to control the proportion of pulse-like signals in each frequency band. In addition to the typical fundamental frequency parameter of the voiced excitation, additional parameters are included which control one or more pulse amplitudes and positions for the pulsed excitation. This model allows additional features of speech and audio signals important for high quality reproduction to be efficiently modeled.

In another aspect, generally, analysis methods are provided for estimating the improved speech model parameters. For pulsed parameter estimation, an error criterion with reduced sensitivity to time shifts is used to reduce computation and improve performance. Pulsed parameter estimation performance is further improved using the estimated voiced strength parameter to reduce the weighting of frequency bands which are strongly voiced when estimating the pulsed parameters.

In another aspect, generally, methods for quantizing the improved speech model parameters are provided. The voiced, unvoiced, and pulsed strength parameters are quantized using a weighted vector quantization method using a novel error criterion for obtaining high quality quantization. The fundamental frequency and pulse position parameters are efficiently quantized based on the quantized strength parameters.

In one general aspect, a method of analyzing a digitized signal to determine model parameters for the digitized signal is provided. The method includes receiving a digitized signal, determining a voiced strength for the digitized signal by evaluating a first function, and determining a pulsed strength for the digitized signal by evaluating a second function. The voiced strength and the pulsed strength may be determined, for example, at regular intervals of time. In some implementations, the voiced strength and the pulsed strength may be determined on one or more frequency bands. In addition, the same function may be used as both the first function and the second function.

The voiced strength and the pulsed strength may be used to encode the digitized signal. In some implementations, the pulse signal may be determined using a pulse signal estimated from the digitized signal. The voiced strength may also be used in determining pulsed strength.

Additionally, the pulsed signal may be determined by combining a transform magnitude with a transform phase computed from a transform magnitude. The transform phase may be near minimum phase. In some implementations, the pulsed strength may be determined using a pulsed signal estimated from a pulse signal and at least one pulse position.

The pulsed strength may be determined by comparing a pulsed signal with the digitized signal. The comparison may be made using an error criterion with reduced sensitivity to time shifts. The error criterion may compute phase differences between frequency samples and may remove the effect of constant phase differences. Additional implementations of the method of analyzing a digitized signal further include quantizing the pulsed strength using a weighted vector quantization, and quantizing the voiced strength using weighted vector quantization. The voiced strength and the pulsed strength may be used to estimate one or more model parameters. Implementations may also include determining the unvoiced strength.

In another general aspect, a method of synthesizing a signal is provided including determining a voiced signal, determining a voiced strength, determining a pulsed signal, determining a pulsed strength, dividing the voiced signal and the pulsed signal into two or more frequency bands, and combining the voiced signal and the pulsed signal based on the voiced strength and the pulsed strength. The pulsed signal may be determined by combining a transform magnitude with a transform phase computed from the transform magnitude.

In another general aspect, a method of synthesizing a signal is provided. The method includes determining a voiced signal; determining a voiced strength; determining a pulsed signal; determining a pulsed strength; determining an unvoiced signal; determining an unvoiced strength; dividing the voiced signal, pulsed signal, and unvoiced signal into two or more frequency bands; and combining the voiced signal, the pulsed signal, and the unvoiced signal based on the voiced strength, the pulsed strength, and the unvoiced strength.

In another general aspect, a method of quantizing speech model parameters is provided. The method includes determining the voiced error between a voiced strength parameter and quantized voiced strength parameters, determining the pulsed error between a pulsed strength parameter and quantized pulsed strength parameters, combining the voiced error and the pulsed error to produce a total error, and selecting the quantized voice strength and the quantized pulsed strength which produce the smallest total error.

In another general aspect, a method of quantizing speech model parameters is provided. The method includes determining a quantized voiced strength, determining a quantized pulsed

strength. The method further includes either quantizing a fundamental frequency based on the quantized voice strength and the quantized pulsed strength or quantizing a pulse position based on the quantized voiced strength and the quantized pulsed strength. The fundamental frequency may be quantized to a constant when the quantized voiced strength is zero for all frequency bands and the pulse position may be quantized to a constant when the quantized voiced strength is nonzero in any frequency band.

The details of one or more implementations are set forth in the accompanying drawings and the description below. Other features and advantages will be apparent from the description and drawings, and from the claims.

Brief Description of the Drawings

Fig. 1 is a block diagram of a speech synthesis system using an improved speech model.

Fig. 2 is a block diagram of an analysis system for estimating parameters of the improved speech model.

Fig. 3 is a block diagram of a pulsed analysis unit that may be used with the analysis system of Fig. 2.

Fig. 4 is a block diagram of a pulsed analysis with reduced complexity.

Fig. 5 is a block diagram of an excitation parameter quantization system.

Detailed Description

Figs. 1-5 show the structure of a system for speech coding, the various blocks and units of which may be implemented with software.

Fig. 1 shows a speech synthesis system 10 that uses an improved speech model which augments the typical excitation parameters with additional parameters for higher quality speech synthesis. Speech synthesis system 10 includes a voiced synthesis unit 11, an unvoiced synthesis unit 12, and a pulsed synthesis unit 13. The signals produced by these units are added together by a summation unit 14.

In addition to parameters which control the proportion of quasi-periodic and noise-like signals in each frequency band, a parameter is added which controls the proportion of pulse-like signals in each frequency band. These parameters are functions of time (t) and frequency (ω) and are denoted by $V(t,\omega)$ for the quasi-periodic voiced strength, $U(t,\omega)$ for the noise-like unvoiced strength, and $P(t,\omega)$ for the pulsed signal strength. Typically, the voiced strength parameter

$V(t,\omega)$ varies between zero indicating no voiced signal at time t and frequency ω and one indicating the signal at time t and frequency ω is entirely voiced. The unvoiced strength and pulse strength parameters behave in a similar manner. Typically, the voiced strength parameters are constrained so that they sum to one (i.e., $V(t,\omega) + U(t,\omega) + P(t,\omega) = 1$).

The voiced strength parameter $V(t,\omega)$ has an associated vector of parameters $\underline{v}(t,\omega)$ which contains voiced excitation parameters and voiced system parameters. The voiced excitation parameters can include a time and frequency dependent fundamental frequency $\omega_0(t,\omega)$ (or equivalently a pitch period $n_0(t,\omega)$). In this implementation, the unvoiced strength parameter $U(t,\omega)$ has an associated vector of parameters $\underline{u}(t,\omega)$ which contains unvoiced excitation parameters and unvoiced system parameters. The unvoiced excitation parameters may include, for example, statistics and energy distribution. Similarly, the pulsed excitation strength parameter $P(t,\omega)$ has an associated vector of parameters $\underline{p}(t,\omega)$ containing pulsed excitation parameters and pulsed system parameters. The pulsed excitation parameters may include one or more pulse positions $t_0(t,\omega)$ and amplitudes.

The voiced parameters $V(t,\omega)$ and $\underline{v}(t,\omega)$ control voiced synthesis unit 11. Voiced synthesis unit 11 synthesizes the quasi-periodic voiced signal using one of several known methods for synthesizing voiced signals. One method for synthesizing voiced signals is disclosed in U.S. Pat. No. 5,195,166, titled "Methods for Generating the Voiced Portion of Speech Signals," which is incorporated by reference. Another method is that used by the MBE vocoder which sums the outputs of sinusoidal oscillators with amplitudes, frequencies, and phases that are interpolated from one frame to the next to prevent discontinuities. The frequencies of these oscillators are set to the harmonics of the fundamental (except for small deviations due to interpolation). In one implementation, the system parameters are samples of the spectral envelope estimated as disclosed in U.S. Pat. No. 5,754,974, titled "Spectral Magnitude Representation for Multi-Band Excitation Speech Coders," which is incorporated by reference. The amplitudes of the harmonics are weighted by the voiced strength $V(t,\omega)$ as in the MBE vocoder. The system phase may be estimated from the samples of the spectral envelope as disclosed in U.S. Pat. No. 5,701,390, titled "Synthesis of MBE-Based Coded Speech using Regenerated Phase Information," which is incorporated by reference.

The unvoiced parameters $U(t,\omega)$ and $\underline{u}(t,\omega)$ control unvoiced synthesis unit 12. Unvoiced synthesis unit 12 synthesizes the noise-like unvoiced signal using one of several known methods for synthesizing unvoiced signals. One method is that used by the MBE vocoder which generates

samples of white noise. These white noise samples are then transformed into the frequency domain by applying a window and fast Fourier transform (FFT). The white noise transform is then multiplied by a noise envelope signal to produce a modified noise transform. The noise envelope signal adjusts the energy around each spectral envelope sample to the desired value. The unvoiced signal is then synthesized by taking the inverse FFT of the modified noise transform, applying a synthesis window, and overlap adding the resulting signals from adjacent frames.

The pulsed parameters $P(t,\omega)$ and $\underline{p}(t,\omega)$ control pulsed synthesis unit 13. Pulsed synthesis unit 13 synthesizes the pulsed signal by synthesizing one or more pulses with the positions and amplitudes contained in $\underline{p}(t,\omega)$ to produce a pulsed excitation signal. The pulsed excitation is then passed through a filter generated from the system parameters. The magnitude of the filter as a function of frequency ω is weighted by the pulsed strength $P(t,\omega)$. Alternatively, the magnitude of the pulses as a function of frequency can be weighted by the pulsed strength.

The voiced signal, unvoiced signal, and pulsed signal produced by units 11, 12, and 13 are added together by summation unit 14 to produce the synthesized speech signal.

Fig. 2 shows a speech analysis system 20 that estimates improved model parameters from an input signal. The speech analysis system 20 includes a sampling unit 21, a voiced analysis unit 22, an unvoiced analysis unit 23, and a pulsed analysis unit 24. The sampling unit 21 samples an analog input signal to produce a speech signal $s_0(n)$. It should be noted that sampling unit 21 operates remotely from the analysis units in many applications. For typical speech coding or recognition applications, the sampling rate ranges between 6 kHz and 16 kHz.

The voiced analysis unit 22 estimates the voiced strength $V(t,\omega)$ and the voiced parameters $\underline{v}(t,\omega)$ from the speech signal $s_0(n)$. The unvoiced analysis unit 23 estimates the unvoiced strength $U(t,\omega)$ and the unvoiced parameters $\underline{u}(t,\omega)$ from the speech signal $s_0(n)$. The pulsed analysis unit 24 estimates the pulsed strength $P(t,\omega)$ and the pulsed signal parameters $\underline{p}(t,\omega)$ from the speech signal $s_0(n)$. The vertical arrows between analysis units 22-24 indicate that information flows between these units to improve parameter estimation performance.

The voiced analysis and unvoiced analysis units can use known methods such as those used for the estimation of MBE model parameters as disclosed in U.S. Pat. No. 5,715,365, titled "Estimation of Excitation Parameters" and U.S. Pat. No. 5,826,222, titled "Estimation of Excitation Parameters," both of which are incorporated by reference. The described implementation of the pulsed analysis unit uses new methods for estimation of the pulsed parameters.

Referring to Fig. 3, the pulsed analysis unit 24 includes a window and Fourier transform unit 31, an estimate pulse FT and synthesize pulsed FT unit 32, and a compare unit 33. The pulsed analysis unit 24 estimates the pulsed strength $P(t,\omega)$ and the pulsed parameters $\underline{p}(t,\omega)$ from the speech signal $s_0(n)$.

The window and Fourier transform unit 31 multiplies the input speech signal $s_0(n)$ by a window $w(t,n)$ centered at time t to obtain a windowed signal $s(t,n)$. The window used is typically a Hamming window or Kaiser window and is typically constant as a function of t so that $w(t,n) = w_0(n-t)$. The length of the window $w(t,n)$ typically ranges between 5 ms and 40 ms. The Fourier transform (FT) of the windowed signal $S(t,\omega)$ is typically computed using a fast Fourier transform (FFT) with a length greater than or equal to the number of samples in the window. When the length of the FFT is greater than the number of windowed samples, the additional samples in the FFT are zeroed.

The estimate pulse FT and synthesize pulsed FT unit 32 estimates a pulse from $S(t,\omega)$ and then synthesizes a pulsed signal transform $\hat{S}(t,\omega)$ from the pulse estimate and a set of pulse positions and amplitudes. The synthesized pulsed transform $\hat{S}(t,\omega)$ is then compared to the speech transform $S(t,\omega)$ using compare unit 33. The comparison is performed using an error criterion. The error criterion can be optimized over the pulse postions, amplitudes, and pulse shape. The optimum pulse positions, amplitudes, and pulse shape become the pulsed signal parameters $\underline{p}(t,\omega)$. The error between the speech transform $S(t,\omega)$ and the optimum pulsed transform $\hat{S}(t,\omega)$ is used to compute the pulsed signal strength $P(t,\omega)$.

A number of techniques exist for estimating the pulse Fourier transform. For example, the pulse can be modeled as the impulse response of an all-pole filter. The coefficients of the all-pole filter can be estimated using well known algorithms such as the autocorrelation method or the covariance method. Once the pulse is estimated, the pulsed Fourier transform can be estimated by adding copies of the pulse with the positions and amplitudes specified. The pulsed Fourier transform is then compared to the speech transform using an error criterion such as weighted squared error. The error criterion is evaluated at all possible pulse positions and amplitudes or some constrained set of positions and amplitudes to determine the best pulse positions, amplitudes, and pulse FT.

Another technique for estimating the pulse Fourier transform is to estimate a minimum phase component from the magnitude of the short time Fourier transform (STFT) $|S(t,\omega)|$ of the speech. This minimum phase component may be combined with the speech transform magnitude

to produce a pulse transform estimate. Other techniques for estimating the pulse Fourier transform include pole-zero models of the pulse and corrections to the minimum phase approach based on models of the glottal pulse shape.

Some implementations emply an error criterion having reduced sensitivity to time shifts (linear phase shifts in the Fourier transform). This type of error criterion can lead to reduced computational requirements since the number of time shifts at which the error criterion needs to be evaluated can be significantly reduced. In addition, reduced sensitivity to linear phase shifts improves robustness to phase distortions which are slowly changing in frequency. These phase distortions are due to the transmission medium or deviations of the actual system from the model. For example, the following equation may be used as an error criterion:

$$E(t) = \min_{\theta} \int_{-\pi}^{\pi} G(t,\omega) \left| S(t,\omega)S^*(t,\omega-\Delta\omega) - e^{j\theta}\hat{S}(t,\omega)\hat{S}^*(t,\omega-\Delta\omega) \right|^2 d\omega \tag{1}$$

In Equation (1), $S(t,\omega)$ is the speech STFT, $\hat{S}(t,\omega)$ is the pulsed transform, $G(t,\omega)$ is a time and frequency dependent weighting, and θ is a variable used to compensate for linear phase offsets. To see how θ compensates for linear phase offsets, it is useful to consider an example. Suppose the speech transform is exactly matched with the pulsed transform except for a linear phase offset so that $\hat{S}(t,\omega) = e^{-j\omega t_0}S(t,\omega)$. Substituting this relation into Equation (1) yields

$$E(t) = \min_{\theta} \int_{-\pi}^{\pi} G(t,\omega) \left| S(t,\omega)S^*(t,\omega-\Delta\omega) \left[1 - e^{j(\theta-\Delta\omega t_0)}\right] \right|^2 d\omega \tag{2}$$

which is minimized over θ at $\theta_{min} = \Delta\omega t_0$. In addition, once θ_{min} is known, the time shift t_0 can be estimated by

$$t_0 = \frac{\theta_{min}}{\Delta\omega} \tag{3}$$

where $\Delta\omega$ is typically chosen to be the frequency interval between adjacent FFT samples.

Equation (1) is minimized by choosing θ as follows

$$\theta_{min}(t) = \arctan\left[\int_{-\pi}^{\pi} G(t,\omega)S(t,\omega)S^*(t,\omega-\Delta\omega)\hat{S}^*(t,\omega)\hat{S}(t,\omega-\Delta\omega)d\omega \right]. \tag{4}$$

When computing $\theta_{min}(t)$ using Equation (4), if $G(t,\omega) = 1$, the frequency weighting is approximately $|S(t,\omega)|^4$. This tends to weight frequency regions with higher energy too heavily relative to frequency regions of lower energy. $G(t,\omega)$ may be used to adjust the frequency weighting. The following function for $G(t,\omega)$ may be used to improve performance in typical applications:

$$G(t,\omega) = \frac{F(t,\omega)}{\sqrt{|S(t,\omega)S^*(t,\omega-\Delta\omega)\hat{S}^*(t,\omega)\hat{S}(t,\omega-\Delta\omega)|}} \tag{5}$$

where $F(t,\omega)$ is a time and frequency weighting function. There are a number of choices for $F(t,\omega)$ which are useful in practice. These include $F(t,\omega) = 1$, which is simple to implement and achieves good results for many applications. A better choice for many applications is to make $F(t,\omega)$ larger in frequency regions with higher pulse-to-noise ratios and smaller in regions with lower pulse-to-noise ratios. In this case, "noise" refers to non-pulse signals such as quasi-periodic or noise-like signals. In one implementation, the weighting $F(t,\omega)$ is reduced in frequency regions where the estimated voiced strength $V(t,\omega)$ is high. In particular, if the voiced strength $V(t,\omega)$ is high enough that the synthesized signal would consist entirely of a voiced signal at time t and frequency ω then $F(t,\omega)$ would have a value of zero. In addition, $F(t,\omega)$ is zeroed out for $\omega < 400$ Hz to avoid deviations from minimum phase typically present at low frequencies. Perceptually based error criteria can also be factored into $F(t,\omega)$ to improve performance in applications where the synthesized signal is eventually presented to the ear.

After computing $\theta_{min}(t)$, a frequency dependent error $E(t,\omega)$ may be defined as:

$$E(t,\omega) = G(t,\omega)\left|S(t,\omega)S_w(t,\omega-\Delta\omega) - e^{j\theta_{min}}\hat{S}(t,\omega)\hat{S}^*(t,\omega-\Delta\omega)\right|^2. \tag{6}$$

The error $E(t,\omega)$ is useful for computation of the pulsed signal strength $P(t,\omega)$. When computing the error $E(t,\omega)$, the weighting function $F(t,\omega)$ is typically set to a constant of one. A small value of $E(t,\omega)$ indicates similarity between the speech transform $S(t,\omega)$ and the pulsed transform $\hat{S}(t,\omega)$, which indicates a relatively high value of the pulsed signal strength $P(t,\omega)$. A large value of $E(t,\omega)$ indicates dissimilarity between the speech transform $S(t,\omega)$ and the pulsed transform $\hat{S}(t,\omega)$, which indicates a relatively low value of the pulsed signal strength $P(t,\omega)$.

Fig. 4 shows a pulsed Analysis unit 24 that includes a window and FT unit 41, a synthesize phase unit 42, and a minimize error unit 43. The pulsed analysis unit 24 estimates the pulsed strength $P(t,\omega)$ and the pulsed parameters from the speech signal $s_0(n)$ using a reduced complexity implementation. The window and FT unit 41 operates in the same manner as previously described for unit 31. In this implementation, the number of pulses is reduced to one per frame in order to reduce computation and the number of parameters. For applications such as speech coding, reduction of the number of parameters is helpful for reduction of speech coding rates. The synthesize phase unit 42 computes the phase of the pulse Fourier transform using well known homomorphic vocoder techniques for computing a Fourier transform with minimum phase

from the magnitude of the speech STFT $|S(t,\omega)|$. The magnitude of the pulse Fourier transform is set to $|S(t,\omega)|$. The system parameter output $\rho(t,\omega)$ consists of the pulse Fourier transform.

The minimize error unit 43 computes the pulse position t_0 using Equations (3) and (4). For this implementation, the pulse position $t_0(t,\omega)$ varies with frame time t but is constant as a function of ω. After computing θ_{min}, the frequency dependent error $E(t,\omega)$ is computed using Equation (6). The normalizing function $D(t,\omega)$ is computed using

$$D(t,\omega) = G(t,\omega)\,|S(t,\omega)S^*(t,\omega-\Delta\omega)|^2 \tag{7}$$

and applied to the computation of the pulsed excitation strength

$$P(t,\omega) = \begin{cases} 0, & P'(t,\omega) < 0 \\ P'(t,\omega), & 0 \le P'(t,\omega) \le 1 \\ 1, & P'(t,\omega) > 1 \end{cases} \tag{8}$$

where

$$P'(t,\omega) = \frac{1}{2}\log_2\left(\frac{2\tau\bar{D}(t,\omega)}{\bar{E}(t,\omega)}\right), \tag{9}$$

$\bar{E}(t,\omega)$ and $\bar{D}(t,\omega)$ are frequency smoothed versions of $E(t,\omega)$ and $D(t,\omega)$, and τ is a threshold typically set to a constant of 0.1. Since $\bar{E}(t,\omega)$ and $\bar{D}(t,\omega)$ are frequency smoothed (low pass filtered), they can be downsampled in frequency without loss of information. In one implementation, $\bar{E}(t,\omega)$ and $\bar{D}(t,\omega)$ are computed for eight frequency bands by summing $E(t,\omega)$ and $D(t,\omega)$ over all ω in a particular frequency band. Typical band edges for these 8 frequency bands for an 8 kHz sampling rate are 0 Hz, 375 Hz, 875 Hz, 1375 Hz, 1875 Hz, 2375 Hz, 2875 Hz, 3375 Hz, and 4000 Hz.

It should be noted that the above frequency domain computations are typically carried out using frequency samples computed using fast Fourier transforms (FFTs). Then, the integrals are computed using summations of these frequency samples.

Referring to Fig. 5, an excitation parameter quantization system 50 includes a voiced/unvoiced/pulsed (V/U/P) strength quantizer unit 51 and a fundamental and pulse position quantizer unit 52. Excitation parameter quantization system 50 jointly quantizes the voiced strength $V(t,\omega)$, the unvoiced strength $U(t,\omega)$, and the pulsed strength $P(t,\omega)$ to produce the quantized voiced strength $\check{V}(t,\omega)$, the quantized unvoiced strength $\check{U}(t,\omega)$, and the quantized pulsed strength $\check{P}(t,\omega)$ using V/U/P strength quantizer unit 51. Fundamental and pulse position

quantizer unit 52 quantizes the fundamental frequency $\omega_0(t,\omega)$ and the pulse position $t_0(t,\omega)$ based on the quantized strength parameters to produce the quantized fundamental frequency $\check{\omega}_0(t,\omega)$ and the quantized pulse position $\check{t}_0(t,\omega)$.

One implementation uses a weighted vector quantizer to jointly quantize the strength parameters from two adjacent frames using 7 bits. The strength parameters are divided into 8 frequency bands. Typical band edges for these 8 frequency bands for an 8 kHz sampling rate are 0 Hz, 375 Hz, 875 Hz, 1375 Hz, 1875 Hz, 2375 Hz, 2875 Hz, 3375 Hz, and 4000 Hz. The codebook for the vector quantizer contains 128 entries consisting of 16 quantized strength parameters for the 8 frequency bands of two adjacent frames. To reduce storage in the codebook, the entries are quantized so that for a particular frequency band a value of zero is used for entirely unvoiced, one is used for entirely voiced, and two is used for entirely pulsed.

For each codebook index m the error is evaluated using

$$E_m = \sum_{n=0}^{1} \sum_{k=0}^{7} \alpha(t_n,\omega_k) E_m(t_n,\omega_k) \tag{10}$$

where

$$E_m(t_n,\omega_k) = \max\left[\left(V(t_n,\omega_k) - \check{V}_m(t_n,\omega_k)\right)^2, \left(1-\check{V}_m(t_n,\omega_k)\right)\left(P(t_n,\omega_k) - \check{P}_m(t_n,\omega_k)\right)^2\right], \tag{11}$$

$\alpha(t_n,\omega_k)$ is a frequency and time dependent weighting typically set to the energy in the speech transform $S(t_n,\omega_k)$ around time t_n and frequency ω_k, $\max(a,b)$ evaluates to the maximum of a or b, and $\check{V}_m(t_n,\omega_k)$ and $\check{P}_m(t_n,\omega_k)$ are the quantized voicing strength and quantized pulsed strength. The error E_m of Equation (10) is computed for each codebook index m and the codebook index is selected which minimizes E_m.

In another preferred embodiment, the error $E_m(t_n,\omega_k)$ of Equation (11) is replaced by

$$E_m(t_n,\omega_k) = \gamma_m(t_n,\omega_k) + \beta\left(1-\check{V}_m(t_n,\omega_k)\right)\left(1-\gamma_m(t_n,\omega_k)\right)\left(P(t_n,\omega_k) - \check{P}_m(t_n,\omega_k)\right)^2, \tag{12}$$

where

$$\gamma_m(t_n,\omega_k) = \left(V(t_n,\omega_k) - \check{V}_m(t_n,\omega_k)\right)^2 \tag{13}$$

and β is typically set to a constant of 0.5.

If the quantized voiced strength $\check{V}(t,\omega)$ is non-zero at any frequency for the two current frames, then the two fundamental frequencies for these frames are jointly quantized using 9 bits, and the pulse positions are quantized to zero (center of window) using no bits.

If the quantized voiced strength $\check{V}(t,\omega)$ is zero at all frequencies for the two current frames and the quantized pulsed strength $\check{P}(t,\omega)$ is non-zero at any frequency for the current two frames, then the two pulse positions for these frames may be quantized using, for example 9 bits, and the fundamental frequencies are set to a value of, for example, 64.84 Hz using no bits.

If the quantized voiced strength $\check{V}(t,\omega)$ and the quantized pulsed strength $\check{P}(t,\omega)$ are both zero at all frequencies for the current two frames, then the two pulse positions for these frames are quantized to zero, and the fundamental frequencies for these frames may be jointly quantized using 9 bits.

Other implementations are within the following claims.

What is claimed is: